

# On Maximal Correlation, Hypercontractivity, and the Data Processing Inequality studied by Erkip and Cover

Venkat Anantharam\*, Amin Gohari†, Sudeep Kamath\*, Chandra Nair‡

\*EECS Department, University of California, Berkeley,  
 {ananth, sudeep}@eecs.berkeley.edu

†EE Department, Sharif University of Technology, Tehran, Iran  
 aminzadeh@sharif.edu

‡ IE Department, The Chinese University of Hong Kong  
 chandra@ie.cuhk.edu.hk

## Abstract

In this paper we provide a new geometric characterization of the Hirschfeld-Gebelein-Rényi maximal correlation of a pair of random  $(X, Y)$ , as well as of the chordal slope of the nontrivial boundary of the hypercontractivity ribbon of  $(X, Y)$  at infinity. The new characterizations lead to simple proofs for some of the known facts about these quantities. We also provide a counterexample to a data processing inequality claimed by Erkip and Cover, and find the correct tight constant for this kind of inequality.

## I. INTRODUCTION

There are various measures available to quantify the dependence between two random variables. A well-known such measure for real-valued random variables is the Pearson correlation coefficient  $\rho_p(X, Y) := \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$ , which quantifies the linear dependence between the two random variables. A closely related measure, called the Hirschfeld-Gebelein-Rényi maximal correlation, or simply the maximal correlation, measures the cosine of the angle between the linear subspaces of mean zero square integrable real-valued random variables defined by the individual random variables, as below.

*Definition 1:* Given random variables  $X$  and  $Y$ , the Hirschfeld-Gebelein-Rényi maximal correlation of  $(X, Y)$  is defined as follows:

$$\rho_m(X; Y) := \max_{(f(X), g(Y)) \in \mathcal{S}} \mathbb{E}[f(X)g(Y)], \quad (1)$$

where  $\mathcal{S}$  is the collection of pairs of real-valued random variables  $f(X)$  and  $g(Y)$  such that

$$\mathbb{E}f(X) = \mathbb{E}g(Y) = 0, \text{ and } \mathbb{E}f^2(X) = \mathbb{E}g^2(Y) = 1.$$

If  $\mathcal{S}$  is empty (which happens precisely when at least one of  $X$  and  $Y$  is constant almost surely) then one defines  $\rho_m(X; Y)$  to be 0.  $\square$

This measure, first introduced by Hirschfeld [9] and Gebelein [6] and then studied by Rényi [17], has found interesting applications in information theory.

As a general remark, to stay clear of technicalities, we restrict ourselves throughout this paper to discrete random variables  $(X, Y)$  taking values in  $\mathcal{X} \times \mathcal{Y}$  with  $|\mathcal{X}|, |\mathcal{Y}| < \infty$ . Further we assume that  $\mathbb{P}(X = x) > 0, \forall x \in \mathcal{X}$  and  $\mathbb{P}(Y = y) > 0, \forall y \in \mathcal{Y}$ . We will use  $:=$  and occasionally  $=:$  for equality by definition.

*Definition 2:* For any real-valued random variable  $X$  and real number  $p \neq 0$ , define  $\|X\|_p := (\mathbb{E}|X|^p)^{\frac{1}{p}}$ . Define  $\|X\|_0 := \exp(\mathbb{E}(\log |X|))$ . For  $p \leq 0$ ,  $\|X\|_p = 0$  if  $\mathbb{P}(|X| = 0) > 0$ .  $\square$

Rényi [17] derived an alternate characterization to  $\rho_m(X, Y)$  as follows:

$$\rho_m(X; Y) = \max_{f(X): \mathbb{E}f(X)=0, \mathbb{E}[f^2(X)]=1} \|\mathbb{E}[f(X)|Y]\|_2. \quad (2)$$

Maximal correlation has interesting connections to the hypercontractivity of Markov operators, as demonstrated by Ahlswede and Gács in [1].

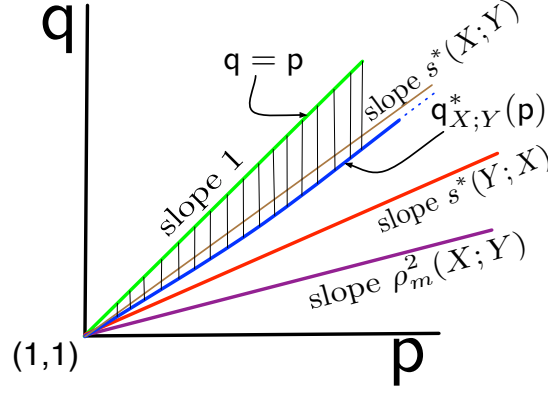


Fig. 1. The blue curve is an illustration of  $q_{X;Y}^*(p)$  (this curve is not convex in general). The brown line represents the ‘chordal’ slope  $\frac{q_{X;Y}^*(p)-1}{p-1}$  as  $p \rightarrow \infty$ , which turns out to be  $s^*(X;Y)$ . The red line is the slope of  $q_{X;Y}^*(p)$  at  $(1,1)$  defined by  $\lim_{p \downarrow 1} \frac{q_{X;Y}^*(p)-1}{p-1}$  and turns out to be  $s^*(Y;X)$ . The purple line passes through  $(1,1)$  and has slope  $\rho_m^2(X;Y)$ .

*Definition 3:* For  $p \geq 1$  define

$$q_{X;Y}^*(p) := \inf \{q : \|\mathbb{E}[g(Y)|X]\|_p \leq \|g(Y)\|_q \ \forall g : \mathcal{Y} \mapsto \mathbb{R}\}. \quad \square$$

*Remark 1:* Ahlswede and Gacs [1] characterize hypercontractivity in terms of  $s_p(X,Y) := \frac{q_{X;Y}^*(p)}{p}$ , for  $p \geq 1$ .

If  $r(x)$  and  $p(x)$  are probability distributions on the same finite set, we write  $D(r(x)||p(x))$  for the relative entropy distance of  $r(x)$  from  $p(x)$ , i.e.

$$D(r(x)||p(x)) := \sum_x r(x) \log \frac{r(x)}{p(x)}.$$

To proceed to discuss the results of this paper, we need the following definition.

*Definition 4:* Let  $X$  and  $Y$  be random variables with joint distribution  $(X,Y) \sim p(x,y)$ . We define

$$s^*(X;Y) := \sup_{r(x) \neq p(x)} \frac{D(r(y)||p(y))}{D(r(x)||p(x))}, \quad (3)$$

where  $r(y)$  denotes the  $y$ -marginal distribution of  $r(x,y) := r(x)p(y|x)$  and the supremum on the right hand side is over all probability distributions  $r(x)$  that are different from the probability distribution  $p(x)$ . If either  $X$  or  $Y$  is a constant, we define  $s^*(X;Y)$  to be 0.  $\square$

*Remark:* From the data processing inequality for relative entropies it is immediate that  $s^*(X;Y) \leq 1$ . Further,  $s^*(X;Y)$  can be regarded as a function of the input distribution  $p(x)$  corresponding to a channel  $p(y|x)$ .

Below, we outline some of the properties of  $q_{X;Y}^*(p)$  combining results from [11] and from Theorems 3 and 5 in [1].

*Theorem 1:* The following statements hold:

- (a) For any fixed  $p > 1$ ,  $q_{X;Y}^*(p) \geq 1$  with equality *if and only if*  $X$  and  $Y$  are independent.
- (b)  $q_{X;Y}^*(1) = 1$  and  $\frac{q_{X;Y}^*(p)}{p}$  is monotonically decreasing in  $p$ .
- (c)  $\frac{q_{X;Y}^*(p)-1}{p-1} \geq \rho_m^2(X;Y)$ .
- (d) The chordal slope of  $q_{X;Y}^*(p)$  at infinity, defined by  $\lim_{p \rightarrow \infty} \frac{q_{X;Y}^*(p)-1}{p-1}$ , exists and is equal to  $s^*(X;Y)$ .
- (e)  $\lim_{p \downarrow 1} \frac{q_{X;Y}^*(p)-1}{p-1} = s^*(Y;X)$ .

*Remark 2:* Hypercontractive inequalities (and their counterpart for  $p < 1$ , called *reverse hypercontractive inequalities*) also play an important role in analysis, probability theory, and discrete Fourier analysis. Interested readers can refer to the introduction in [16] for a brief summary of their development and impact in these areas. For results and applications of hypercontractivity and reverse hypercontractivity in information theory, interested readers can refer to [11].

In this paper we will provide alternate characterizations of both  $\rho_m^2(X, Y)$  and  $s^*(X; Y)$ . Fix a channel  $p(y|x)$ , fix  $\lambda \in [0, 1]$ , and consider the function<sup>1</sup> of the probability distribution of  $X$  denoted by  $t_\lambda(X)$  which is defined by

$$t_\lambda(X) := H(Y) - \lambda H(X).$$

We will show in Theorem 4 that  $\rho_m^2(X, Y)$  is the smallest  $\lambda$  such that  $t_\lambda(X)$  has a positive semidefinite Hessian at  $p(x)$  and  $s^*(X; Y)$  is the smallest  $\lambda$  such that  $t_\lambda(X)$  matches its lower convex envelope, denoted by  $\mathcal{K}[t_\lambda](X)$ , at  $p(x)$ .

In [4, Theorem 8] it was claimed that the following inequality holds:

$$I(U; Y) \leq \rho_m^2(X; Y)I(U; X), \quad \forall U - X - Y.$$

It turns out that this inequality is incorrect; we will provide a counter example in this paper. Further we will show (Theorem 4) that the following inequality holds, with a tight constant:

$$I(U; Y) \leq s^*(X; Y)I(U; X), \quad \forall U - X - Y.$$

The error in the proof in [4, Theorem 8] seems to be a subtle, yet significant one. A similar error has also occurred in [10], where the authors independently rediscover the erroneous result of [4, Theorem 8] using similar techniques.

#### A. Alternate characterizations of the Hirschfeld-Gebelein-Rényi maximal correlation

In this section we will review some alternate characterizations of the Hirschfeld-Gebelein-Rényi maximal correlation which are known in the literature.

1) *Rényi's characterization:* As mentioned earlier, Rényi derived the following “one-function” alternate characterization for  $\rho_m(X; Y)$  [17]:

$$\rho_m^2(X; Y) = \max_{f(X): \mathbb{E}f(X)=0, \mathbb{E}[f^2(X)]=1} \mathbb{E}[\mathbb{E}[f(X)|Y]^2]. \quad (4)$$

The validity of this characterization can be proved by fixing  $f$  with  $\mathbb{E}[f(X)] = 0$  and  $\mathbb{E}[f^2(X)] = 1$  and showing that setting  $g(Y) = \alpha \mathbb{E}[f(X)|Y]$  maximizes  $\mathbb{E}[f(X)g(Y)]$  among all functions  $g$  with  $\mathbb{E}[g(Y)] = 0$  and  $\mathbb{E}[g^2(Y)] = 1$  when  $\alpha \geq 1$  is chosen so that  $\alpha^2 \mathbb{E}(\mathbb{E}[f(X)|Y]^2) = 1$ . This is a simple consequence of the Cauchy-Schwartz inequality.

2) *Distribution simulation characterization:* Consider a random variable  $X'$  such that  $X - Y - X'$  is Markov and  $(X, Y) \stackrel{d}{=} (X', Y)$ . Then

$$\rho_m^2(X; Y) = \max_{f(X): \mathbb{E}f(X)=0, \mathbb{E}[f^2(X)]=1} \mathbb{E}[f(X)f(X')]. \quad (5)$$

This result follows from Rényi's characterization which was given in (4) above. Since  $(X, Y) \stackrel{d}{=} (X', Y)$ , we have  $\mathbb{E}[f(X)|Y] = \mathbb{E}[f(X')|Y]$ . Hence  $\mathbb{E}[\mathbb{E}[f(X)|Y]^2] = \mathbb{E}[\mathbb{E}[f(X)|Y]\mathbb{E}[f(X')|Y]] \stackrel{(a)}{=} \mathbb{E}[\mathbb{E}[f(X)f(X')|Y]] = \mathbb{E}[f(X)f(X')]$ , where (a) holds because  $X - Y - X'$ .

3) *Singular value characterization:* For finite valued random variables maximal correlation  $\rho_m(X; Y)$  can also be characterized [18] by the second largest singular value of the matrix  $Q$  with entries  $Q_{x,y} = \frac{p(x,y)}{\sqrt{p(x)p(y)}}$ . This result can be seen by writing  $\mathbb{E}[f(X)g(Y)]$  as  $\sum_{x,y} (f(x)\sqrt{p(x)})Q(x,y)(g(y)\sqrt{p(y)})$ , observing that  $\sum_x \sqrt{p(x)}Q(x,y) = \sqrt{p(y)}$  and  $\sum_y Q(x,y)\sqrt{p(y)} = \sqrt{p(x)}$ , and that the conditions  $\mathbb{E}[f(X)] = 0$  and  $\mathbb{E}[g(Y)] = 0$  are respectively equivalent to requiring that  $x \mapsto f(x)\sqrt{p(x)}$  is orthogonal to  $x \mapsto \sqrt{p(x)}$  and that  $y \mapsto g(y)\sqrt{p(y)}$  is orthogonal to  $y \mapsto \sqrt{p(y)}$ .

There is a simple formula for  $\rho_m(X; Y)$  if at least one of  $X$  or  $Y$  is binary-valued, which is most easily seen by using the singular value characterization:

$$\rho_m^2(X; Y) = \left[ \sum_{x,y} \frac{p(x,y)^2}{p(x)p(y)} \right] - 1. \quad (6)$$

<sup>1</sup>We abuse notation when we write  $t_\lambda(X)$ . We really wish to think of  $t_\lambda$  as a function of the probability distribution of  $X$ .

This follows from observing that  $\rho_m^2(X; Y)$  is the second largest eigenvalue of both  $QQ^T$  and  $Q^TQ$ . If one of these is a 2 by 2 matrix, we can find the second largest eigenvalue by computing the trace and subtracting the largest eigenvalue, i.e. 1, from it.

### B. Properties of $\rho_m(X; Y)$

In this section, we will present some known properties of the maximal correlation  $\rho_m(X; Y)$ .

1) *Tensorization of  $\rho_m(X; Y)$* : The following theorem shows that maximal correlation *tensorizes*. It was proved by Witsenhausen in [18]. For a function of probability distributions to have the property of the first sentence of the theorem is what it means to say that it tensorizes.

*Theorem 2:* (Witsenhausen [18]) If  $(X_1, Y_1), (X_2, Y_2)$  are independent, then

$$\rho_m(X_1, X_2; Y_1, Y_2) = \max\{\rho_m(X_1; Y_1), \rho_m(X_2; Y_2)\}.$$

In particular if  $(X_1, Y_1), (X_2, Y_2)$  are i.i.d., then  $\rho_m(X_1, X_2; Y_1, Y_2) = \rho_m(X_1; Y_1)$ .  $\square$

The elegant proof in [14] (for finite valued random variables) uses the singular value characterization and is reproduced below. When  $(X_1, Y_1)$  is independent of  $(X_2, Y_2)$  it is immediate that the matrix  $Q$  defined by  $Q_{x_1, x_2, y_1, y_2} = \frac{p_1(x_1, y_1)p_2(x_2, y_2)}{\sqrt{p_1(x_1)p_2(x_2)p_1(y_1)p_2(y_2)}}$  is the Kronecker product of the corresponding individual matrices  $\hat{Q}_{x_1, y_1} = \frac{p_1(x_1, y_1)}{\sqrt{p_1(x_1)p_1(y_1)}}$  and  $\tilde{Q}_{x_2, y_2} = \frac{p_2(x_2, y_2)}{\sqrt{p_2(x_2)p_2(y_2)}}$ , i.e.  $Q = \hat{Q} \otimes \tilde{Q}$ . It is known that the singular values of  $Q$  are given as the set of products of one singular value of  $\hat{Q}$  with one singular value of  $\tilde{Q}$ . Since the largest singular values of each of the three matrices is unity, it is immediate that the second largest singular value of  $Q$  is  $\max\{\rho_m(X_1; Y_1), \rho_m(X_2; Y_2)\}$ .

Witsenhausen [18] showed that the maximal correlation of two random variables gives the answer to the following problem: consider two agents, the first of whom observes  $X^n$ , while the second observes  $Y^n$ , where  $(X_i, Y_i), 1 \leq i \leq n$ , are i.i.d. copies of  $(X, Y)$ . Each agent makes a binary decision based on the sequence available to it. The entropy of the each binary decision should be bounded away from zero by a constant. Witsenhausen showed that the probability of agreement between these decisions can be made to converge to 1, as  $n$  converges to infinity, if and only if  $\rho_m(X; Y) = 1$ . This is a version of the main result in the path-breaking work of Gács and Körner [5], which introduced the concept of Gács-Körner common information.

Erkip and Cover [4] studied the problem of investment in the stock market with side information of limited rate with the aim of quantifying the value of the side information in improving the growth rate of wealth. In one part of their much broader contribution, they present a data processing inequality which claims that

$$I(U; Y) \leq \rho_m^2(X; Y)I(U; X), \quad \forall U - X - Y$$

where  $\rho_m(X; Y)$  is the Hirschfeld-Gebelein-Rényi maximal correlation between the random variables  $X$  and  $Y$ . As we stated earlier, this inequality is incorrect.

Kang and Ulukus illustrated some applications of maximal correlation in distributed source and channel coding problems [12]. Beigi has introduced a quantum version of the maximal correlation for bipartite quantum states, and has shown that this measure fully characterizes bipartite states from which common randomness distillation under local operations is possible [2].

Recently Kamath and Anantharam [11] have used maximal correlation to study the problem of non-interactive simulation of joint distributions. They also used hypercontractivity and reverse hypercontractivity to show that under certain conditions these can provide stronger impossibility results for the simulation problem than those obtained by maximal correlation.

### C. Alternate characterization and properties of $q_{X;Y}^*(p)$

In [11], the authors defined the following region which can be used to characterize  $q_{X;Y}^*(p)$ .

*Definition 5:* For a pair of random variables  $(X, Y) \sim p(x, y)$  on  $\mathcal{X} \times \mathcal{Y}$ , the *hypercontractivity ribbon* is the subset

$$\mathcal{R}_{X;Y} \subseteq \{(p, q) \in \mathbb{R}^2 : 1 \leq q \leq p \text{ or } 1 \geq q \geq p\}$$

defined by<sup>2</sup>

<sup>2</sup>This characterization of the hypercontractivity ribbon is given in [11]. Another characterization, which is closer to how hypercontractivity is normally discussed in the literature, will be mentioned later.

- $(1, 1) \in \mathcal{R}_{X;Y}$ ;
- For  $1 \leq q \leq p$ ,  $(p, q) \in \mathcal{R}_{X;Y}$  iff

$$\mathbb{E}f(X)g(Y) \leq \|f(X)\|_{p'}\|g(Y)\|_q \quad \forall f : \mathcal{X} \mapsto \mathbb{R}, g : \mathcal{Y} \mapsto \mathbb{R}; \quad (7)$$

- For  $1 \geq q \geq p$ ,  $(p, q) \in \mathcal{R}_{X;Y}$  iff

$$\mathbb{E}f(X)g(Y) \geq \|f(X)\|_{p'}\|g(Y)\|_q \quad \forall f : \mathcal{X} \mapsto (0, \infty), g : \mathcal{Y} \mapsto (0, \infty). \quad (8)$$

When  $1 \leq q < p$ , inequalities such as (7) are referred to in the literature as *hypercontractive* inequalities and when  $1 \geq q > p$ , inequalities such as (8) are referred to as *reverse hypercontractive* inequalities.  $\square$

Then one can alternatively define  $q_{X;Y}^*(p)$  according to

$$q_{X;Y}^*(p) := \inf\{q : (q, p) \in \mathcal{R}_{X;Y}\}, p \geq 1.$$

The equivalence of this characterization to that in definition 3 is proved in [11]. The proof is similar to that of Renyi's alternate characterization of  $\rho_m(X, Y)$  and is a straightforward application of Hölder's inequality.

Likewise, we can define  $\mathcal{R}_{Y;X}$ . In general,  $\mathcal{R}_{X;Y} \neq \mathcal{R}_{Y;X}$ , but the two are related by an intimate duality relationship that is clear from (7) and (8):

$$(p, q) \in \mathcal{R}_{X;Y} \iff (q', p') \in \mathcal{R}_{Y;X}.$$

Using this duality relationship [11] establishes that  $\lim_{p \rightarrow 1} \frac{q_{X;Y}^*(p)-1}{p-1} = \frac{d}{dp} q_{X;Y}^*(p) \Big|_{p=1} = s^*(Y; X)$ .

*Remark 3:* In general,  $s^*(X; Y) \neq s^*(Y; X)$  as shown by the following example. Let  $(X, Y)$  be 0-1 valued with  $\mathbb{P}(X = 0) = 0.85, \mathbb{P}(Y = 0) = 0.39, \mathbb{P}(X = Y = 0) = 0.36$ . Then, computation gives us  $s^*(X; Y) = 0.045\dots, s^*(Y; X) = 0.029\dots$

Most of the applications of hypercontractivity traces its roots to the following *tensorization* property of the hypercontractive ribbon.

*Theorem 3:* ([1], [16]) If  $(X_1, Y_1)$  and  $(X_2, Y_2)$  are independent, then  $\mathcal{R}_{(X_1, X_2);(Y_1, Y_2)} = \mathcal{R}_{X_1;Y_1} \cap \mathcal{R}_{X_2;Y_2}$ . In particular, if  $(X_1, Y_1)$  and  $(X_2, Y_2)$  are i.i.d., then  $\mathcal{R}_{(X_1, X_2);(Y_1, Y_2)} = \mathcal{R}_{X_1;Y_1}$ .  $\square$

Theorem 3 can be thought of as saying that the whole hypercontractivity ribbon tensorizes, since it says that for each  $(p, q)$  we have

$$\mathbb{1}((p, q) \notin \mathcal{R}_{(X_1, X_2);(Y_1, Y_2)}) = \max\{\mathbb{1}((p, q) \notin \mathcal{R}_{X_1;Y_1}), \mathbb{1}((p, q) \notin \mathcal{R}_{X_2;Y_2})\}.$$

A consequence of this then is that  $s^*(X; Y)$  tensorizes, i.e. for  $(X_1, Y_1)$  and  $(X_2, Y_2)$  independent,

$$s^*(X_1, X_2; Y_1, Y_2) = \max\{s^*(X_1; Y_1), s^*(X_2; Y_2)\}.$$

We will give a alternate proof of the tensorization of  $s^*(X; Y)$  later using our new characterization involving the function  $t_\lambda(X)$  that was introduced earlier. A direct proof of this tensorization can be obtained as follows. The direction  $s^*(X_1, X_2; Y_1, Y_2) \geq \max\{s^*(X_1; Y_1), s^*(X_2; Y_2)\}$  is immediate; hence we only show the non-trivial direction. Note that for any  $r(x_1, x_2) \neq p(x_1, x_2)$  we have

$$\begin{aligned} \frac{D(r(y_1, y_2)||p(y_1, y_2))}{D(r(x_1, x_2)||p(x_1, x_2))} &\stackrel{(a)}{=} \frac{D(r(y_1)||p(y_1)) + \sum_{y_1} r(y_1)D(r(y_2|y_1)||p(y_2))}{D(r(x_1)||p(x_1)) + \sum_{x_1} r(x_1)D(r(x_2|x_1)||p(x_2))} \\ &= \frac{D(r(y_1)||p(y_1)) + \sum_{y_1} r(y_1)D(\sum_{x_1} r(x_1|y_1)r(y_2|x_1)||p(y_2))}{D(r(x_1)||p(x_1)) + \sum_{x_1} r(x_1)D(r(x_2|x_1)||p(x_2))} \\ &\stackrel{(b)}{\leq} \frac{D(r(y_1)||p(y_1)) + \sum_{y_1} \sum_{x_1} r(x_1|y_1)r(y_1)D(r(y_2|x_1)||p(y_2))}{D(r(x_1)||p(x_1)) + \sum_{x_1} r(x_1)D(r(x_2|x_1)||p(x_2))} \\ &= \frac{D(r(y_1)||p(y_1)) + \sum_{x_1} r(x_1)D(r(y_2|x_1)||p(y_2))}{D(r(x_1)||p(x_1)) + \sum_{x_1} r(x_1)D(r(x_2|x_1)||p(x_2))} \\ &\leq \max\{s^*(X_1; Y_1), s^*(X_2; Y_2)\}. \end{aligned}$$

In the above (a) uses the fact that  $p(x_1 y_1, x_2, y_1) = p_1(x_1, y_1)p_2(x_2, y_2)$  and (b) uses the convexity of  $D(p||q)$  in  $p$ . The last inequality follows from the definition of  $s^*(X_1; Y_1), s^*(X_2; Y_2)$ , and our assumption that  $r(x_1, x_2) \neq p(x_1, x_2)$  which guarantees that at least one of the terms in the denominator is non-zero. Finally taking sup over all such  $r(x_1, x_2)$  we obtain the non-trivial direction  $s^*(X_1, X_2; Y_1, Y_2) \leq \max\{s^*(X_1; Y_1), s^*(X_2; Y_2)\}$ .

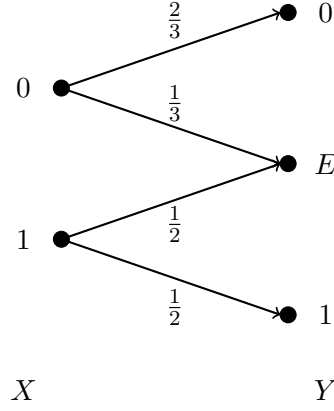


Fig. 2. An asymmetric erasure channel.

## II. MAIN RESULTS

One of the main contributions of this paper is a correction to the data processing inequality claimed by Erkip and Cover in [4, Theorem 8]. We provide a counterexample to their claim and point out a location in their proof where the argument is incomplete. We then find the correct constant to get a tight data processing inequality of the type they considered.

### A. Counterexample to the Erkip-Cover data processing inequality

In [4, Theorem 8], Erkip and Cover claimed that

$$I(U; Y) \leq \rho_m^2(X; Y) I(U; X) \quad (9)$$

holds whenever  $U - X - Y$  form a Markov chain. Furthermore they claimed that,  $\rho_m^2(X; Y)$  is the minimum such constant, i.e.

$$\sup_{U: U-X-Y, I(U; X) > 0} \frac{I(U; Y)}{I(U; X)} = \rho_m^2(X; Y). \quad (10)$$

We will first provide a counterexample to these claims and then point where there is a gap in their argument. In a subsequent subsection we will identify  $s^*(X; Y)$  as the correct constant to replace  $\rho_m^2(X; Y)$  in (9) and (10).

1) *Counterexample to (9) and (10)*: Let  $X$  be a binary random variable with  $p(X = 0) = \frac{1}{2}$ . Define  $p(x, y)$  by passing  $X$  through the asymmetric erasure channel given in Fig. 2. Using Equation (6), one can verify for this pair  $(X, Y)$  that  $\rho_m^2(X; Y) = 0.6$ .

Suppose we construct  $U$  satisfying  $U - X - Y$  such that  $U|X = 0 \sim \text{Ber}(0.1)$ ,  $U|X = 1 \sim \text{Ber}(0.4)$ . Then  $I(U; Y) = 0.055770\dots$  and  $I(U; X) = 0.09130\dots$ , so that  $\frac{I(U; Y)}{I(U; X)} = 0.6108\dots > 0.6 = \rho_m^2(X; Y)$ , and this contradicts (10).

It can be shown in a reasonably straightforward manner, using our characterization in Theorem 4, that  $s^*(X; Y) = \frac{1}{2} \log_2 \left( \frac{12}{5} \right) = 0.631517\dots$  for this pair of random variables  $(X, Y)$ . Simulation shows that for a suitable sequence of  $U_i$  with  $I(U_i; X) \rightarrow 0$ , we can have  $\frac{I(U_i; Y)}{I(U_i; X)}$  approach  $s^*(X; Y)$  for this example. A sequence of such  $U_i$  is shown in the table below.

$P(U = 1 X = 0)$	$P(U = 1 X = 1)$	$I(U; Y)$	$I(U; X)$	$\frac{I(U; Y)}{I(U; X)}$
0.1	0.4	0.055770...	0.09130...	0.6108...
0.01	0.23	0.062321...	0.099958...	0.6234...
0.001	0.102	0.031038...	0.049379...	0.6285...
0.0001	0.04	0.012507...	0.019838...	0.6304...
0.00001	0.01474	0.0046418...	0.0073545...	0.6311...
0.000001	0.005232	0.0016507...	0.0026145...	0.6313...
0.0000001	0.0018146	0.00057285...	0.00090716...	0.6314...
0.00000001	0.00061973	0.000195672...	0.000309852...	0.63150...

The error of the Erkip-Cover proof seems to lie in their use of a Taylor's series expansion. Consider the expansion in the left column of page 1037 of their paper [4], where they use their equation (16) to expand around  $p(\tilde{v})$ . It is possible that  $p(\tilde{v})$  is zero for some  $\tilde{v}$  and this causes an error as the derivative in this direction is infinity and the Taylor's series expansion is no longer valid. As our counterexample shows this seems to be a significant but subtle error that cannot be worked around.

Some of the works that use this incorrect result of [4], such as [3] and [19], are affected by this error. A claim similar to that of [4], which appears in [10], is also false.<sup>3</sup>

### B. A geometric characterization of $\rho_m^2(X; Y)$ and $s^*(X; Y)$

Given  $p(x, y)$ , we can treat  $p(y|x)$  as a channel, and then consider the function of the input distribution  $p(x)$ , defined by

$$t_\lambda(X) := H(Y) - \lambda H(X),$$

where  $\lambda$  is a constant in  $[0, 1]$ . Observe that the function is concave when  $\lambda = 0$  and convex when  $\lambda = 1$ .<sup>4</sup>

We write  $\mathcal{K}[t_\lambda](X)$  for the convex hull of  $t_\lambda(X)$ . If  $\mathcal{K}[t_\lambda](X) = t_\lambda(X)$  at  $p(x)$  for some  $\lambda$ , then note that for any  $\lambda_1 \geq \lambda$

$$\begin{aligned} \mathcal{K}[t_{\lambda_1}](X) &= \mathcal{K}[t_\lambda - (\lambda_1 - \lambda)H](X) \\ &\geq \mathcal{K}[t_\lambda](X) - (\lambda_1 - \lambda)H(X). \end{aligned}$$

Here the inequality comes from  $\mathcal{K}[f + g] \geq \mathcal{K}[f] + \mathcal{K}[g]$  and since  $-(\lambda_1 - \lambda)H(X)$  is convex. Therefore at  $p(x)$  we will have that

$$t_{\lambda_1}(X) \geq \mathcal{K}[t_{\lambda_1}](X) \geq \mathcal{K}[t_\lambda](X) - (\lambda_1 - \lambda)H(X) = t_\lambda(X) - (\lambda_1 - \lambda)H(X) = t_{\lambda_1}(X).$$

Thus we see that if  $\mathcal{K}[t_\lambda](X) = t_\lambda(X)$  at  $p(x)$  for some  $\lambda$  then  $\mathcal{K}[t_{\lambda_1}](X) = t_{\lambda_1}(X)$  at  $p(x)$  for all  $\lambda_1 \geq \lambda$ .

The following theorem gives a geometric interpretation of  $\rho_m^2(X; Y)$  and  $s^*(X; Y)$  in terms of the behaviour of the function  $t_\lambda(X)$  and identifies  $s^*(X; Y)$  as the correct replacement for  $\rho_m^2(X; Y)$  in (9) and (10).

*Theorem 4:* The following statements hold:

- 1)  $\rho_m^2(X; Y)$  is the minimum value of  $\lambda$  such that the function  $t_\lambda(X)$  has a positive semidefinite Hessian at  $p(x)$ .
- 2)  $s^*(X; Y)$  is the minimum value of  $\lambda$  such that the function  $t_\lambda(X)$  touches its lower convex envelope at  $p(x)$ , i.e. such that  $\mathcal{K}[t_\lambda](X) = t_\lambda(X)$  at  $p(x)$ . Furthermore,

$$\sup_{U: U-X-Y, I(U; X) > 0} \frac{I(U; Y)}{I(U; X)} = s^*(X; Y).$$

□

*Proof of 1):* This follows from Rényi's characterization of the maximal correlation, given in (4) above. Take an arbitrary multiplicative perturbation of the form  $p_\epsilon(x) = p(x)(1 + \epsilon f(x))$ . For  $p_\epsilon$  to stay a valid perturbation we need  $\mathbb{E}[f(X)] = 0$ . Furthermore we can normalize  $f$  by assuming that  $\mathbb{E}[f^2(X)] = 1$ . The second derivative in  $\epsilon$  of  $H(Y) - \lambda H(X)$  is equal to [8]

$$-\mathbb{E}[\mathbb{E}[f(X)|Y]^2] + \lambda \mathbb{E}[f^2(X)] = -\mathbb{E}[\mathbb{E}[f(X)|Y]^2] + \lambda,$$

which is non-negative as long as  $\lambda \geq \mathbb{E}[\mathbb{E}[f(X)|Y]^2]$ . Thus the minimum value  $\lambda^*$  such that the second derivative is non-negative for all local perturbations is

$$\lambda^* = \max_{f(X): \mathbb{E}[f(X)] = 0, \mathbb{E}[f^2(X)] = 1} \mathbb{E}[\mathbb{E}[f(X)|Y]^2] = \rho_m^2(X; Y),$$

where the last equality follows from Rényi's characterization of maximal correlation. ■

<sup>3</sup>This paper studies the ratio  $\frac{I(U; Y)}{I(U; X)}$  when  $I(U; X)$  is very small. However, as pointed out in [4], the supremum of  $\frac{I(U; Y)}{I(U; X)}$  occurs when  $I(U; X) \rightarrow 0$ . So the problem studied by [10] is the same as that of [4].

<sup>4</sup>This convexity at  $\lambda = 1$  follows from the fact that for any  $U - X - Y$  we have  $I(U; X) \geq I(U; Y)$  or equivalently  $H(Y) - H(X) \leq H(Y|U) - H(X|U)$ .

*Proof of 2):* Consider the minimum value of  $\lambda$ , say  $\lambda^\dagger$ , such that the function  $t_\lambda(X)$  touches its lower convex envelope at  $p(x)$ . Thus, equivalently we are looking for the minimum  $\lambda$  such that for  $(X, Y) \sim p(x, y)$  we have

$$H(Y) - \lambda H(X) \leq H(Y|U) - \lambda H(X|U), \quad \forall U : U - X - Y.$$

Note that if  $U$  is independent of  $X$ , i.e.  $I(U; X) = 0$  then the above inequality is always true. Equivalently we require the minimum  $\lambda$  such that,

$$\lambda \geq \frac{I(U; Y)}{I(U; X)}, \quad \forall U : U - X - Y \text{ with } I(U; X) > 0.$$

Thus,

$$\lambda^\dagger = \sup_{U: U-X-Y, I(U; X) > 0} \frac{I(U; Y)}{I(U; X)}.$$

*Remark:* Since  $t_\lambda(X) = \mathcal{K}[t_\lambda](X)$  at  $p(x)$  implies that the Hessian of  $t_\lambda(X)$  at  $p(x)$  is positive semidefinite, we have

$$\sup_{U: U-X-Y, I(U; X) > 0} \frac{I(U; Y)}{I(U; X)} \geq \rho_m^2(X; Y). \quad (11)$$

It remains to show that  $\lambda^\dagger = s^*(X; Y)$  or equivalently that

$$\sup_{U: U-X-Y, I(U; X) > 0} \frac{I(U; Y)}{I(U; X)} = s^*(X; Y).$$

From standard cardinality bounding arguments, it suffices to consider  $|\mathcal{U}| \leq |\mathcal{X}| + 1$  to determine the value of  $\sup_{U: U-X-Y, I(U; X) > 0} \frac{I(U; Y)}{I(U; X)}$ .

For any  $|\mathcal{U}| \leq |\mathcal{X}| + 1$  and  $U - X - Y$  a Markov chain with  $I(U; X) > 0$  and  $X \sim p(x)$ , denote  $P(U = u) =: w_u$ ,  $P(X = x|U = u) =: r_u(x)$ . Clearly  $\sum_u w_u r_u(x) = p(x)$ . Let the channel-induced distributions on  $Y$  corresponding to the  $r_u(x)$  be denoted by  $r_u(y)$  respectively. Then elementary manipulations yield

$$\frac{I(U; Y)}{I(U; X)} = \frac{\sum_{u \in \mathcal{U} : r_u(x) \neq p(x)} w_u D(r_u(y) \| p(y))}{\sum_{u \in \mathcal{U} : r_u(x) \neq p(x)} w_u D(r_u(x) \| p(x))} \leq \sup_{r(x) \neq p(x)} \frac{D(r(y) \| p(y))}{D(r(x) \| p(x))},$$

where  $r(y)$  denotes the channel-induced probability distribution on  $Y$  corresponding to the probability distribution  $r(x)$  on  $X$ .

Since the above holds for all  $U$  such that  $U - X - Y$  is a Markov chain and  $I(U; X) > 0$ , we have

$$\sup_{U: U-X-Y, I(U; X) > 0} \frac{I(U; Y)}{I(U; X)} \leq \sup_{r(x)} \frac{D(r(y) \| p(y))}{D(r(x) \| p(x))} = s^*(X; Y),$$

where the last equality is by definition, see (3) above.

To show the other direction, we assume that  $s^*(X; Y) > 0$ , else there is nothing to prove. Let  $\delta \in (0, s^*(X; Y))$  be arbitrary. We also assume without loss of generality that  $p(x) > 0 \forall x \in \mathcal{X}$  and  $p(y) > 0 \forall y \in \mathcal{Y}$ , since otherwise we could have simply changed the definition of  $\mathcal{X}$  and  $\mathcal{Y}$ .

Let  $\mathcal{U}_\epsilon := \{1, 2\}$ . Fix a sufficiently small  $\epsilon > 0$  and define  $U_\epsilon$  by:

- $w_1 = \epsilon, r_1(x) = r^*(x)$ ,
- $w_2 = 1 - \epsilon, r_2(x) = p(x) + \frac{\epsilon}{1-\epsilon}(p(x) - r^*(x)) = \frac{1}{1-\epsilon}p(x) - \frac{\epsilon}{1-\epsilon}r^*(x)$ ,

where  $r^*(x) \neq p(x)$  is a probability distribution satisfying  $\frac{D(r^*(y) \| p(y))}{D(r^*(x) \| p(x))} > s^*(X; Y) - \delta$ . For sufficiently small  $\epsilon > 0$ ,

we will have that  $r_2(x)$  is a probability distribution. Note that  $w_1 + w_2 = 1$  and  $w_1 r_1(x) + w_2 r_2(x) = p(x) \forall x \in \mathcal{X}$ . Clearly  $I(U_\epsilon; Y) > 0$ , since  $I(X; Y) = 0$  would have implied that  $s^*(X; Y) = 0$ .

For any  $0 < \lambda < s^*(X; Y) - \delta$  define the function

$$g(\epsilon) := I(U_\epsilon; Y) - \lambda I(U_\epsilon; X).$$

<sup>5</sup>Indeed our proof below indicates that even a binary  $U$  suffices.

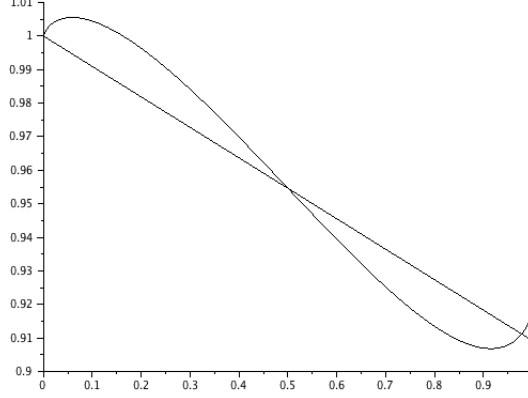


Fig. 3. Plot of  $p(x) \mapsto H(Y) - 0.6H(X)$  for the asymmetric erasure channel given in Fig. 2. The X-axis is  $P(X = 0)$ . The straight line is drawn to connect the value of the curve at  $P(X = 0) = 0$  to that at  $P(X = 0) = \frac{1}{2}$ , to visually demonstrate that this line is not tangent to the curve at  $P(X = 0) = \frac{1}{2}$ .

We have

$$\begin{aligned} \frac{dg(\epsilon)}{d\epsilon} &= -\frac{d}{d\epsilon} \left( \epsilon H(r^*(y)) + (1-\epsilon)H\left(\frac{1}{1-\epsilon}p(y) - \frac{\epsilon}{1-\epsilon}r^*(y)\right) \right) \\ &\quad + \lambda \frac{d}{d\epsilon} \left( \epsilon H(r^*(x)) + (1-\epsilon)H\left(\frac{1}{1-\epsilon}p(x) - \frac{\epsilon}{1-\epsilon}r^*(x)\right) \right) \\ &= -H(r^*(y)) + H\left(\frac{p(y) - \epsilon r^*(y)}{1-\epsilon}\right) + \lambda H(r^*(x)) - \lambda H\left(\frac{p(x) - \epsilon r^*(x)}{1-\epsilon}\right) \\ &\quad - \sum_y \frac{r^*(y) - p(y)}{1-\epsilon} \log\left(\frac{p(y) - \epsilon r^*(y)}{1-\epsilon}\right) + \lambda \sum_x \frac{r^*(x) - p(x)}{1-\epsilon} \log\left(\frac{p(x) - \epsilon r^*(x)}{1-\epsilon}\right). \end{aligned}$$

Thus

$$\left. \frac{dg(\epsilon)}{d\epsilon} \right|_{\epsilon=0} = D(r^*(y) \| p(y)) - \lambda D(r^*(x) \| p(x)) > 0,$$

where the last inequality is because  $0 < \lambda < s^*(X; Y) - \delta$  and  $\frac{D(r^*(y) \| p(y))}{D(r^*(x) \| p(x))} > s^*(X; Y) - \delta$ . Since  $g(0) = 0$  this implies that for some  $\epsilon' > 0$  we have  $I(U_{\epsilon'}; Y) - \lambda I(U_{\epsilon'}; X) > 0$  or that

$$\sup_{U: U-X-Y, I(U; X) > 0} \frac{I(U; Y)}{I(U; X)} \geq \frac{I(U_{\epsilon'}; Y)}{I(U_{\epsilon'}; X)} > \lambda.$$

Since the above holds for all  $\lambda < s^*(X; Y) - \delta$  we have

$$\sup_{U: U-X-Y, I(U; X) > 0} \frac{I(U; Y)}{I(U; X)} \geq s^*(X; Y) - \delta.$$

Finally, since  $\delta > 0$  is arbitrary, we are done. ■

*Remarks:*

- Note that  $\rho_m^2(X; Y)$  is symmetric in the pair  $(X, Y)$  but  $s^*(X; Y)$  is not, i.e.  $s^*(X; Y) \neq s^*(Y; X)$  in general. Thus,  $\sup_{U: U-X-Y, I(U; X) > 0} \frac{I(U; Y)}{I(U; X)} \neq \sup_{V: X-Y-V, I(V; Y) > 0} \frac{I(V; X)}{I(V; Y)}$  in general, which is a qualitatively different phenomenon than predicted by the incorrect Erkip-Cover claim in (10) above.
- This theorem also explains the motivation for our counterexample of the previous subsection. The plot of  $p(x) \mapsto H(Y) - 0.6H(X)$  for the channel  $p(y|x)$  described earlier is given in Fig. II-B. The second derivative of the function  $p(x) \mapsto H(Y) - 0.6H(X)$  at  $P(X = 0) = \frac{1}{2}$  is zero. This validates the fact that  $\rho_m^2(X; Y) = 0.6$ . It is clear that the lower convex envelope of the curve does not pass through  $P(X = 0) = \frac{1}{2}$ . The straight

line in the figure connects the values of the curve at 0 and  $\frac{1}{2}$  and clearly demonstrates that the line is not a tangent to the curve. Thus it is clear in the figure that  $\rho_m^2(X; Y)$  is the local-convexity condition and not the condition for being on the convex envelope.

- Thm. 8 of [1] asserts that for fixed  $p(y|x)$ ,

$$\max_{p(x)} \rho_m^2(X; Y) = \max_{p(x)} s^*(X; Y).$$

Using the interpretation from Theorem 4 this is immediate since having a positive semidefinite Hessian at all points in the domain implies the graph is convex. Thus, both quantities above equal the minimum value of  $\lambda$  such that the function  $p(x) \mapsto H(Y) - \lambda H(X)$  is convex.

- The above characterization of  $s^*(X; Y)$  is also partly motivated from Körner and Marton's characterization in [13] of less noisy broadcast channels, where they show that for a broadcast channel  $X \rightarrow (Y, Z)$  the following holds:

$$I(U; Y) \geq I(U; Z) \quad \forall U \rightarrow X \rightarrow (Y, Z) \iff D(r(z)||p(z)) \leq D(r(y)||p(y)) \quad \forall r(x), p(x),$$

where  $r(y), r(z)$  are the corresponding channel-induced distributions at  $Y$  and  $Z$  when  $X \sim r(x)$  and similarly  $p(y), p(z)$  are the corresponding channel-induced distributions at  $Y$  and  $Z$  when  $X \sim p(x)$ .

### C. Alternate proof for the tensorization of $s^*(X; Y)$

The above characterization of  $s^*(X; Y)$  results in an alternate proof of its tensorization. This proof is directly motivated by the factorization inequalities in broadcast channels, some of which can be found in [7]. Take a distribution of the form  $p(x_1, x_2, y_1, y_2) = p_1(x_1)p_1(y_1|x_1)p_2(x_2)p_2(y_2|x_2)$ . The easy direction is that  $s^*(X_1X_2; Y_1Y_2) \geq \max(s^*(X_1; Y_1), s^*(X_2; Y_2))$ . This easily follows from the definition of  $s^*(X; Y)$ . Thus the non-trivial part is to show that  $s^*(X_1X_2; Y_1Y_2) \leq \max(s^*(X_1; Y_1), s^*(X_2; Y_2))$ .

Let  $\lambda := \max(s^*(X_1; Y_1), s^*(X_2; Y_2))$ . With  $\mathcal{K}$  denoting the lower convex envelope operator, as earlier, we have  $t_\lambda(X_1) = \mathcal{K}[t_\lambda](X_1)$  at  $p_1(x_1)$  and  $t_\lambda(X_2) = \mathcal{K}[t_\lambda](X_2)$  at  $p_2(x_2)$ , where  $t_\lambda(X_1)$  denotes  $H(Y_1) - \lambda H(X_1)$  and  $t_\lambda(X_2)$  denotes  $H(Y_2) - \lambda H(X_2)$ .

We need to show that  $t_\lambda(X_1, X_2) = \mathcal{K}[t_\lambda](X_1, X_2)$  at  $p_1(x_1)p_2(x_2)$ , where  $t_\lambda(X_1, X_2)$  denotes  $H(Y_1, Y_2) - \lambda H(X_1, X_2)$ , thought of as a function of  $p(x_1, x_2)$ , with the channel given by  $p(y_1, y_2|x_1, x_2) = p_1(y_1|x_1)p_2(y_2|x_2)$ .

Since for any  $W$  satisfying the Markov chain  $W - X_1X_2 - Y_1Y_2$ , we have

$$\begin{aligned} H(Y_1, Y_2|W) - \lambda H(X_1, X_2|W) &= H(Y_1|W) - \lambda H(X_1|W) + H(Y_2|W, Y_1) - \lambda H(X_2|W, X_1) \\ &\geq H(Y_1|W) - \lambda H(X_1|W) + H(Y_2|W, Y_1, X_1) - \lambda H(X_2|W, X_1) \\ &= H(Y_1|W) - \lambda H(X_1|W) + H(Y_2|W, X_1) - \lambda H(X_2|W, X_1), \end{aligned}$$

we conclude that

$$\mathcal{K}[t_\lambda](X_1, X_2) \geq \mathcal{K}[t_\lambda](X_1) + \mathcal{K}[t_\lambda](X_2).$$

This inequality in fact holds for all  $\lambda$  and for all  $p(x_1, x_2)$ , not just for the specific  $\lambda$  under consideration and at  $p_1(x_1)p_2(x_2)$ , which is where we want to use it.

Now with  $(X_1, X_2) \sim p_1(x_1)p_2(x_2)$ , we also have

$$H(Y_1, Y_2) - \lambda H(X_1, X_2) = H(Y_1) - \lambda H(X_1) + H(Y_2) - \lambda H(X_2),$$

i.e. we have  $t_\lambda(X_1, X_2) = t_\lambda(X_1) + t_\lambda(X_2)$  at  $p_1(x_1)p_2(x_2)$ . We can put together the facts so far to write

$$t_\lambda(X_1, X_2) = t_\lambda(X_1) + t_\lambda(X_2) = \mathcal{K}[t_\lambda](X_1) + \mathcal{K}[t_\lambda](X_2) \leq \mathcal{K}[t_\lambda](X_1, X_2),$$

holding for the specific  $\lambda$  as defined above and for  $(X_1, X_2) \sim p_1(x_1)p_2(x_2)$ . But by our characterization of  $s^*(X_1X_2; Y_1Y_2)$ , this implies that  $s^*(X_1X_2; Y_1Y_2) \leq \max\{s^*(X_1; Y_1), s^*(X_2; Y_2)\}$ , completing the proof of the non-trivial direction.

### III. CONCLUSION

In this paper we presented a new geometric characterization of the maximal correlation,  $\rho_m(X; Y)$ , of a pair of discrete random variables  $(X, Y)$  taking values in finite sets. We also presented a new geometric characterization of the chordal slope of the nontrivial boundary of the hypercontractivity ribbon of  $(X, Y)$  at infinity,  $s^*(X; Y)$ . We showed the application of these new characterizations in recovering some of the known results about these quantities in a simple way. We also made a correction to a data processing inequality claimed by Erkip and Cover [4], the error in which has had some knock-on effects in the literature. It would be interesting to find other connections between the curve  $t_\lambda(X)$  that we have associated to the channel  $p(y|x)$  and the entire hypercontractivity ribbon of  $(X, Y)$ , as we vary  $p(x)$ .

### ACKNOWLEDGEMENTS

S. Kamath and V. Anantharam gratefully acknowledge research support from the ARO MURI grant W911NF-08-1-0233, “Tools for the Analysis and Design of Complex Multi-Scale Networks”, from the NSF grant CNS-0910702, and from the NSF Science & Technology Center grant CCF-0939370, “Science of Information”. The work of Chandra Nair was partially supported by the following: an area of excellence grant (Project No. AoE/E-02/08) and two GRF grants (Project Nos. 415810 and 415612) from the University Grants Committee of the Hong Kong Special Administrative Region, China.

### REFERENCES

- [1] R. Ahlswede and P. Gács “Spreading of Sets in Product Spaces and Hypercontraction of the Markov Operator,”
- [2] S. Beigi, “A New Quantum Data Processing Inequality,” arXiv: 1210.1689.
- [3] T. A. Courtade, “Outer Bounds for Multiterminal Source Coding based on Maximal Correlation”, arXiv: 1302.3492
- [4] E. Erkip and T. Cover, “The efficiency of investment information,” IEEE Transactions On Information Theory, vol. 44, pp. 1026-1040, May 1998.
- [5] P. Gács and J. Körner, “Common information is far less than mutual information,” Problems of Control and Information Theory, Vol. 2, No. 2, 1973, pp. 149 -162.
- [6] H. Gebelein, “Das statistische Problem der Korrelation als Variations- und Eigenwert-problem und sein Zusammenhang mit der Ausgleichsrechnung,” Zeitschrift für angew. Math. und Mech. 21, pp. 364-379 (1941).
- [7] Y. Geng and C. Nair, “The capacity region of the two-receiver vector gaussian broadcast channel with private and common messages,” Feb. 2012, 1202.0097.
- [8] A. Gohari and V. Anantharam, “Evaluation of Marton’s Inner Bound for the General Broadcast Channel,” IEEE Transactions on Information Theory, Volume 58, Issue 2, Feb. 2012.
- [9] H. O. Hirschfeld, “A connection between correlation and contingency,” Proc. Cambridge Philosophical Soc. 31, pp 520-524 (1935).
- [10] S.-L. Huang and L. Zheng, “Linear Information Coupling Problems”, IEEE Symposium On Information Theory (ISIT), 2012.
- [11] S. Kamath and V. Anantharam, “Non-interactive Simulation of Joint Distributions: The Hirschfeld-Gebelein-Rnyi Maximal Correlation and the Hypercontractivity Ribbon,” Proceedings of the 50th Annual Allerton Conference on Communications, Control and Computing 2012, Monticello, Illinois.
- [12] W. Kang and S. Ulukus, “A New Data Processing Inequality and Its Applications in Distributed Source and Channel Coding,” IEEE Transactions on Information Theory 57, 56-69 (2011)
- [13] Janos Körner and Katalin Marton, “Comparison of two noisy channels,” Topics in Inform. Theory(ed. by I. Csiszar and P.Elias), Keszthely, Hungary, August, 1975, pp 411-423.
- [14] G. Kumar, “On sequences of pairs of dependent random variables: A simpler proof of the main result using SVD,” On webpage, July 2010, [http://www.stanford.edu/~gowthamr/research/Witsenhausen\\_simpleproof.pdf](http://www.stanford.edu/~gowthamr/research/Witsenhausen_simpleproof.pdf)
- [15] G. Kumar, “Binary Rényi Correlation: A simpler proof of Witsenhausen’s result and a tight lower bound,” On webpage, July 2010, [http://www.stanford.edu/~gowthamr/research/binary\\_renyi\\_correlation.pdf](http://www.stanford.edu/~gowthamr/research/binary_renyi_correlation.pdf)
- [16] E. Mossel, K. Oleszkiewicz and A. Sen, “On Reverse Hypercontractivity”, Geometric and Functional Analysis, 2013, pp. 1-36.
- [17] A. Rényi, “On measures of dependence,” Acta Math. Hung., vol. 10, pp. 441-451, 1959.
- [18] H.S. Witsenhausen, “On sequences of pairs of dependent random variables,” SIAM Journal on Applied Mathematics, vol. 28, no. 1, pp. 100-113, January 1975.
- [19] L. Zhao and Y.-K. Chia, “The efficiency of common randomness generation”, 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2011, pp. 944 - 950.